DOCUMENT RESUME

ED 318 802 TM 014 936

AUTHOR Linacre, John M.
TITLE Rank Ordering or Judge-Awarded Ratings?
PUB DATE Apr 90
NOTE 10p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Comparative Analysis; *Error of Measurement; Essay Tests; Evaluators; *Goodness of Fit; *Item Response Theory; Licensing Examinations (Professions); Mathematical Models; *Quality Control; Rating Scales; *Scoring
IDENTIFIERS *Rank Order; *Rasch Model

ABSTRACT

Rank ordering examinees is an easier task for judges than is awarding numerical ratings. A measurement model for rankings based on Rasch's objectivity axioms provides linear, sample-independent and judge-independent measures. Estimates of examinee measures are obtained from the data set of rankings, along with standard errors and fit statistics. Judge quality-control fit statistics are also obtained for each ordering. An example is provided comparing rating and ranking of an essay examination. Even though rank ordering does not perform measurement as precisely as does a highly discriminating rating scale, measures obtained from rankings can be comparable to those obtained from rating scales commonly in use in certifying situations. The simplification of judge training and the easing of the burden on the judge resulting from the use of rank ordering rather than rating scales suggest that this technique merits close scrutiny by examination boards that currently rely on judges to rate examinee performance. Five tables and three figures are included. (Author/TJH)

Rank Ordering or Judge-Awarded Ratings ?

by

John M. Linacre

MESA Psychometric Laboratory
Department of Education
University of Chicago

Paper presented at
American Educational Research Association Annual Meeting
Boston, Massachusetts
April 1990.

Abstract

Rank ordering examinees is an easier task for judges than awarding numerical ratings. A measurement model for rankings based on Rasch's objectivity axioms provides linear, sample-independent and judge-independent measures. Estimates of examinee measures are obtained from the data set of rankings, along with standard errors and fit statistics. Judge quality-control fit statistics are also obtained for each ordering. An example is provided comparing rating and ranking of an essay examination.

Key words: Rating Scale; Rasch Measurement; Rank order

Introduction: Rating or Ranking?

It is said to be easier for a judge to place examinee performances into an order of merit than to assign numerical ratings (Harper and Misra, 1976). Rating scales require considerable effort in terms of scale design and judge training, and they also require considerable effort on the part of the judges to maintain content-specific standards over a long judging session.

The replacement of numerical rating by rank ordering would simplify the judging task, and, could also, for certain types of examination, increase the precision of the examinee measurements. Ranking removes problems due to judge leniency and to variation task difficulty.

Measurements from Rating Scale Observations

The conversion of rating scale observations into linear, objective measures is well understood (Rasch 1961, Andrich 1978, Wright and Masters 1982), and is an extension of the measurement approach widely applied to multiple-choice tests. Expressed algebraically, the model is:

$$\log (P_{nij}/P_{nij-1}) = B_n - D_i - F_j \qquad \text{for } j=1,J \qquad (1)$$

where

- P_{nij} is the probability of an observation in category j
- P_{nij-1} is the probability of an observation in category j-1
- B_n is the ability of person n
- D_i is the severity of judge i
- F_j is the step difficulty or threshold between categories j-1 and j, where the categories are numbered, say, 0,J, and, for the purposes of this discussion, all judges employ the same category structure.

When ratings for an examinee are available from several judges, these can be used to construct linear measures. Each examinee receives an ability estimate, each judge a severity estimate, and parameters specifying the structure of the rating scale are estimated. Standard errors, which are indicative of reliability, and fit statistics, which are indicative of internal validity, are also obtained.

Measurements from Rank Order Observations

The potential for the use of rank ordering as a measurement device was noted by Thurstone and Chave (1929), and a number of analytical models have been devised (David 1988 Chap. 6). The model employed here is based on the axioms of objective measurement (Rasch 1960, 1980). A rank ordering is conceived to be a set of paired comparisons constrained by the overall ordering imposed on them. The constraint that rank ordering imposes on the paired comparisons is illustrated in Figure 1, in which two of the possible eight sets of paired comparisons of three examinees can be seen to produce inconsistent orderings. Without this constraint, the log-odds of one examinee being ranked higher than another is just the difference between their measures on a linear measurement scale:

$$\log(P_{mn}/P_{nm}) = B_m - B_n \tag{2}$$

where

P_{mn} is the probability that person m out-performs person n,
P_{nm} is the probability that person n out-performs person m,
B_m is the measure of person m in logits (log-odds units),
B_n is the measure of person n.

With the constraint of rank ordering, however, a mathematically determinable adjustment must be made to eliminate the effect of the constraint from the measures (Linacre 1989). The resulting measures are linear and locally independent of the nature of the examinee sample, or the judge population. In this framework, tied rankings present no problem.

In order to estimate the odds of one examinee being ranked higher than another, several independent (i.e. made by different judges) rankings which include each examinee must be obtained. There is no requirement, however, that all rankings include all examinees, or that all judges rank all examinees. It is only necessary that the rankings contain an overlapping network of examinees.

Analytical considerations indicate that asking a judge to rank about 10 examinees at a time is the most computationally efficient. In practice, this number would be determined by the nature of the examination.

The entire data set containing a number of overlapping rankings of the examinees, compiled by several judges acting independently, is subjected to maximum likelihood estimation. The output of the analysis is, for each examinee, a measure, a standard error and a fit statistic (which indicates the quality of the measure obtained), and also, for each judge's ordering, a quality-control fit statistic, which indicates the degree to which that rank ordering is consistent with the overall consensus.

If the examination is attempting to determine performance relative to a criterion, then a performance of known measure, relative to the criterion can be introduced into some of the rankings, and ranked along with the examinees. This will place the examinee performances on a scale marked by the level of the known criterion. The introduction of two known, but different, performances into the ranking scheme would enable the linear scale for this judging session to be equated to a previously established linear scale.

A comparative example

Table 1 contains the ratings collected by Hartog and Rhodes (1936 p.121). Experienced, trained judges were asked to grade essays written for the "Special Place" examination. The subset of ratings analyzed here were marks that were awarded "by impression". 15 judges independently marked 9 essays on a scale from 0 to 100. This marking scheme is more detailed than many used currently, and so in Table 2 the ratings have been reduced to a 0 to 3 point scale, by collapsing 0–29 to 0, 30–39 to 1, 40–49 to 2, 50–100 to 3.

The information in Table 1 was also used to rank order the essays according tc judge, and the ranks are listed in Table 3.

Tables 1,2 and 3 are arranged in order by raw score on Table 1, both for essay and for judge. The best essay is listed first as number 1, and the most lenient judge is listed first as judge 1. Comparing Tables 1 and 2 shows that collapsing the categories has only slightly altered the raw score order of the essays, exchanging essays 7 and 8. The raw score order of the judges is considerably affected, indicating that they may use the original scale slightly idiosyncratically, but it is clear that many of them are similar in severity.

Inspection of Table 3 reveals several interesting features. Judge severity has now disappeared, since the sum of the rankings in each ordering is identical. If measures of judge severity are required, then analysis of the data in Table 1, using each essay to rank the judges in order of severity, yields judge severity estimates. For this data set, the ranked order of essays is the same as that based on the original ratings, which is desirable, but not crucial. Comparing the ordering produced by each judge with the overall ordering, either by eye or by the fit statistics in Table 5, no judge exactly agreed with the consensus, but judge 1, the most lenient judge, and judge 15, the most severe judge, are both close to it.

Table 4 gives the measures for the essays, and Table 5 gives the calibrations for the judges according to the three scoring methods. The more precise the information, the smaller the standard errors, so that it can be seen that the rank order scoring lies between the very detailed 101-point rating scale and the aggregated 4-point scale.

The different discriminations of the scoring methods are reflected in the measures as can be seen by the different slopes in Figure 2. The reliability coefficients for all three methods are larger than .9. In Figure 3, the differences virtually disappear when the measure distributions for the three methods are standardized.

The fit statistics for all three methods, reported in Tables 4 and 5, agree on the worst fitting essay, 3, and judge, 12. Judge 2 has used a narrow range of the scale to rate the essays, leading to mean square fit statistics of less than 1 in Table 5. The order of essays within that range is somewhat in conflict with the consensus, leading to a rank mean square fit statistic greater than 1. The arbitrary way in which the 101-point scale was grouped into a 4-point scale caused an apparent restriction in the range of Judge 6's ratings in Table 2, and an anomalous low mean square fit statistic for Judge 6 on the 4-point scale in Table 5.

Conclusions and importance

This study shows that even though rank ordering does not perform measurement as precisely as a highly discriminating rating scale, measures obtained from rankings can be comparable with those obtained from rating scales commonly in use in certifying situations. The simplification of judge training and the easing of the burden on the judge, resulting from the use of rank ordering rather than rating scales, suggest that this technique merits close scrutiny by examination boards which currently rel, on judges to rate examinee performances.

Bibliography

Andrich, D. (1978) A rating formulation for ordered response categories. Psychometrika 43(4): 561-573.

David, H.A. (1988) The method of paired comparisons. London: Charles Griffin.

Harper, A.E. and Misra, V.S. (1976) Research on Examinations in India. New Delhi: NCERT.

Hartog, P. and Rhodes, E.C. (1936) The marks of examiners. London: MacMillan.

Linacre, J.M. (1989) Many-faceted Rasch Measurement. Chicago: MESA Press.

Rasch, G (1960, 1980) Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Rasch, G (1961) On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 321-333.

Thurstone, L.L. and Chave, E.J. (1929) The measurement of attitude. Chicago: University of Chicago Press.

Wright, B.D. and Masters, G.N. (1982) Rating Scale Analysis. Chicago: MESA Press.

Essay	Judge (using 101 Point Scale) 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Raw Score
1	58	51	53	60	53	45	48	57	58	50	46	48	53	36	46	762
2	52	43	53	48	52	52	61	63	52	39	40	39	50	36	38	718
3	51	47	46	62	37	49	53	53	41	50	38	27	48	48	35	685
4	49	46	51	41	55	46	43	40	43	45	40	33	36	35	29	632
5	45	48	48	40	41	39	47	42	47	38	42	42	32	33	32	616
6	46	50	48	44	43	42	34	30	31	41	43	38	31	33	31	585
7	36	45	39	39	45	41	41	37	36	38	35	38	25	27	21	543
8	43	42	32	29	37	41	37	41	31	34	34	38	32	32	26	529
9	44	37	38	38	37	40	29	27	43	24	29	30	24	30	26	496
	424	409	408	401	400	395	393	390	382	359	347	333	331	310	284	

Table 1. Ratings awarded to 9 essays by 15 judges on 0–100 point scale (Hartog & Rhodes 1936).

Essay	Judge (transformed to 4 Point Scale) 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Score
1	3	3	3	3	3	2	2	3	3	3	2	2	3	1	2	38
2	3	2	3	2	3	3	3	3	3	1	2	1	3	1	1	34
3	3	2	2	3	1	2	3	3	2	3	1	0	2	2	1	30
4	2	2	3	2	3	2	2	2	2	2	2	1	1	1	0	27
5	2	2	2	2	2	1	2	2	2	1	2	2	1	1	1	25
6	2	3	2	2	2	2	1	1	1	2	2	1	1	1	1	24
7	1	2	1	1	2	2	2	1	1	1	1	1	0	0	0	16
8	2	2	1	0	1	2	1	2	1	1	1	1	1	1	0	17
9	2	1	1	1	1	2	0	0	2	0	0	1	0	1	0	12
	20	19	18	16	18	18	16	17	17	14	13	10	12	9	6	

Table 2. Ratings awarded to 9 essays by 15 judges on 0–100 point scale, converted to 0–3 point scale.

Essay	Judge (transformed to Ranks) 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Rank Sum
1	1	1	1½	2	2	4	3	2	1	1½	1	1	1	2½	1	25½
2	2	7	1½	3	3	1	1	1	2	5	4½	3	2	2½	2	40½
3	3	4	6	1	8	2	2	3	6	1½	6	9	3	1	3	58½
4	4	5	3	5	1	3	5	6	4½	3	4½	7	4	4	6	65
5	6	3	4½	6	6	9	4	4	3	6½	3	2	5½	5½	4	72
6	5	2	4½	4	5	5	8	8	8½	4	2	5	7	5½	5	78½
7	9	6	7	7	4	6½	6	7	7	6½	7	5	8	9	9	104
8	8	8	9	9	8	6½	7	5	8½	8	8	5	5½	7	7½	110
9	7	9	8	8	8	8	9	9	4½	9	9	8	9	8	7½	121
	45	45	45	45	45	45	45	45	45	45	45	45	45	45	45	

Table 3. Rank order of essays by judge, based on ratings awarded to 9 essays by 15 judges on 0–100 point scale.

Essay	101 Point Scale MEASURE	ERROR	MNSQ	4 Point Scale MEASURE	ERROR	MNSQ	Rank Order MEASURE	ERROR	MNSQ
1	.52	.06	1.0	3.30	.52	1.0	1.11	.27	.9
2	.38	.06	1.5	2.32	.46	1.2	.61	.18	1.2
3	.29	.06	2.4	1.50	.44	1.9	.25	.14	1.9
4	.12	.06	.6	.91	.43	.6	.15	.14	.6
5	.08	.06	.5	.52	.43	.6	.04	.14	.9
6	−.02	.06	1.3	.33	.43	.8	−.06	.14	1.0
7	−.15	.06	.7	−1.26	.45	.8	−.50	.16	.7
8	−.21	.06	.9	−1.05	.44	.8	−.63	.18	.8
9	−.32	.06	1.2	−2.15	.48	1.5	−.96	.23	1.0
Mean	.08	.06	1.1	.49	.45	1.0	.00	.18	1.0
S.D.	.28	.00	.6	1.76	.03	.5	.60	.04	.4

Table 4. Essay calibrations, standard errors, and fit statistics.

JUDGE	101 Point Scale CALIBRTN	ERROR	MNSQ	4 Point Scale CALIBRTN	ERROR	MNSQ	Rank MNSQ
1	-.29	.07	.5	-1.83	.61	.6	.4
2	-.21	.07	.9	-1.45	.60	.8	1.6
3	-.21	.07	.7	-1.09	.59	.5	.5
4	-.14	.07	1.3	-.38	.58	1.0	1.7
5	-.16	.07	1.4	-1.09	.59	1.2	1.4
6	-.14	.07	.7	-1.09	.59	1.3	.7
7	-.11	.07	1.2	-.38	.58	1.3	.8
8	-.09	.07	1.8	-.74	.59	1.1	.9
9	-.06	.07	1.5	-.74	.59	.8	1.5
10	.05	.07	.9	.29	.57	1.4	.8
11	.12	.07	.7	.63	.57	.7	1.0
12	.20	.08	1.9	1.66	.59	1.5	2.2
13	.20	.08	1.4	.97	.58	1.0	.5
14	.34	.08	1.2	2.02	.60	1.1	.6
15	.50	.09	.6	3.22	.65	.6	.4
Mean	.00	.08	1.1	.49	.45	1.0	1.0
S.D.	.23	.00	.5	1.76	.03	.5	.5

Table 5. Judge calibrations, standard errors and fit statistics.

Paired Comparisons			Representation as rank order
(m,n)	(m,c)	(n,c)	(m,n,c)
(m,n)	(m,c)	(c,n)	(m,c,n)
(m,n)	(c,m)	(n,c)	inconsistent
(m,n)	(c,m)	(c,n)	(c,m,n)
(n,m)	(m,c)	(n,c)	(n,m,c)
(n,m)	(m,c)	(c,n)	inconsistent
(n,m)	(c,m)	(n,c)	(n,c,m)
(n,m)	(c,m)	(c,n)	(c,n,m)

Figure 1. The paired comparison of three examinees, m, n, c, and the equivalent rank orderings. (x,y) indicates that x compares favorably with y.

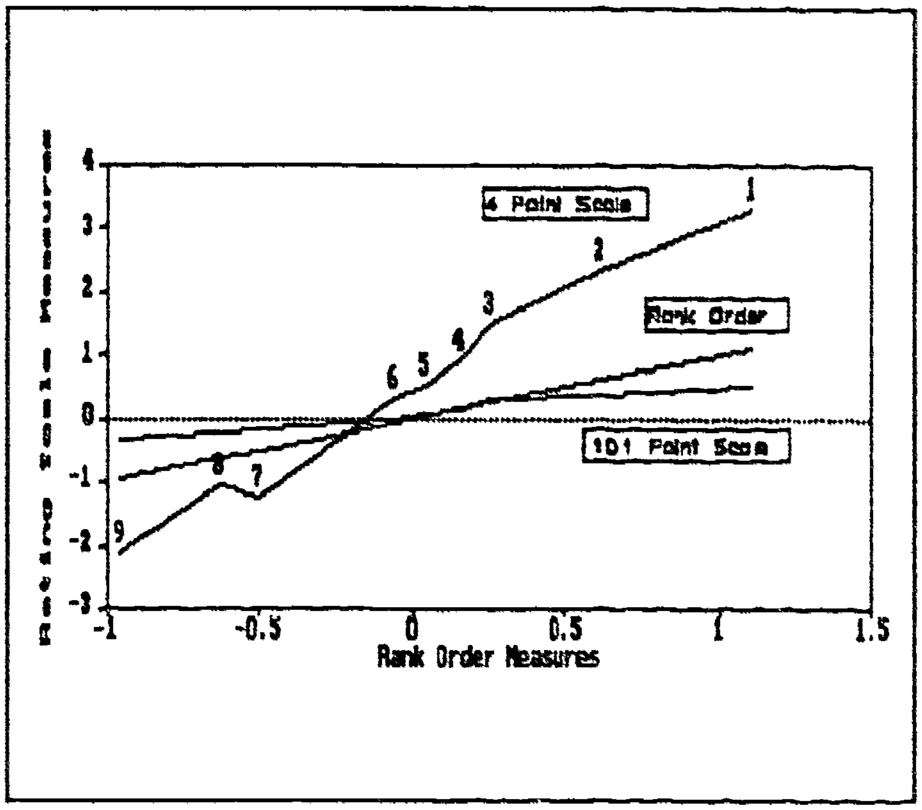

Figure 2. Comparison of essay measures obtained by the three methods. The local zero for each method is that of the mean judge severity, and the slope of the line reflects the discrimination of the scoring method.

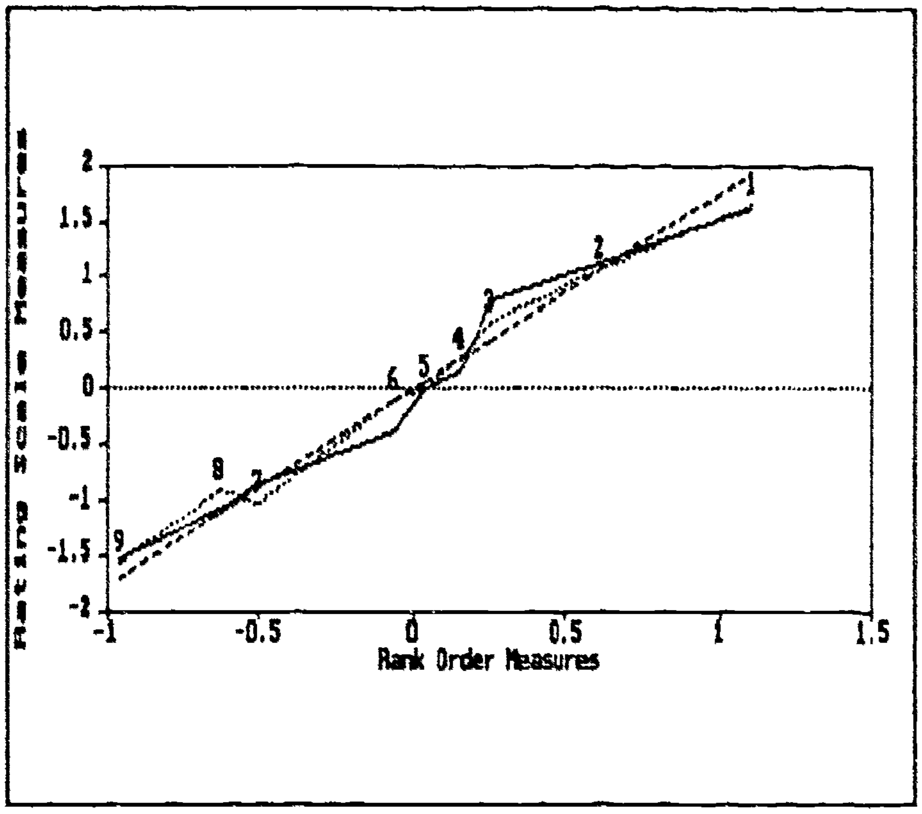

Figure 3. Comparison of essay measures obtained by the three methods. The measures have been equated by setting the mean of the measures for each method to zero, and the standard deviation to 1.